Task 3 – Web Community Sensing

REACTION Meeting – July 12th, 2011

Overview

- Modeling the credibility and authority of news sources and **opinion makers in social networks**

  Twitter data crawling, pre-processing and curation

- Identifying **influential individuals and experts** on a given news topic

  Implementation of a baseline influence model

- Monitoring the **community reaction to news** stories and the polarity of opinions

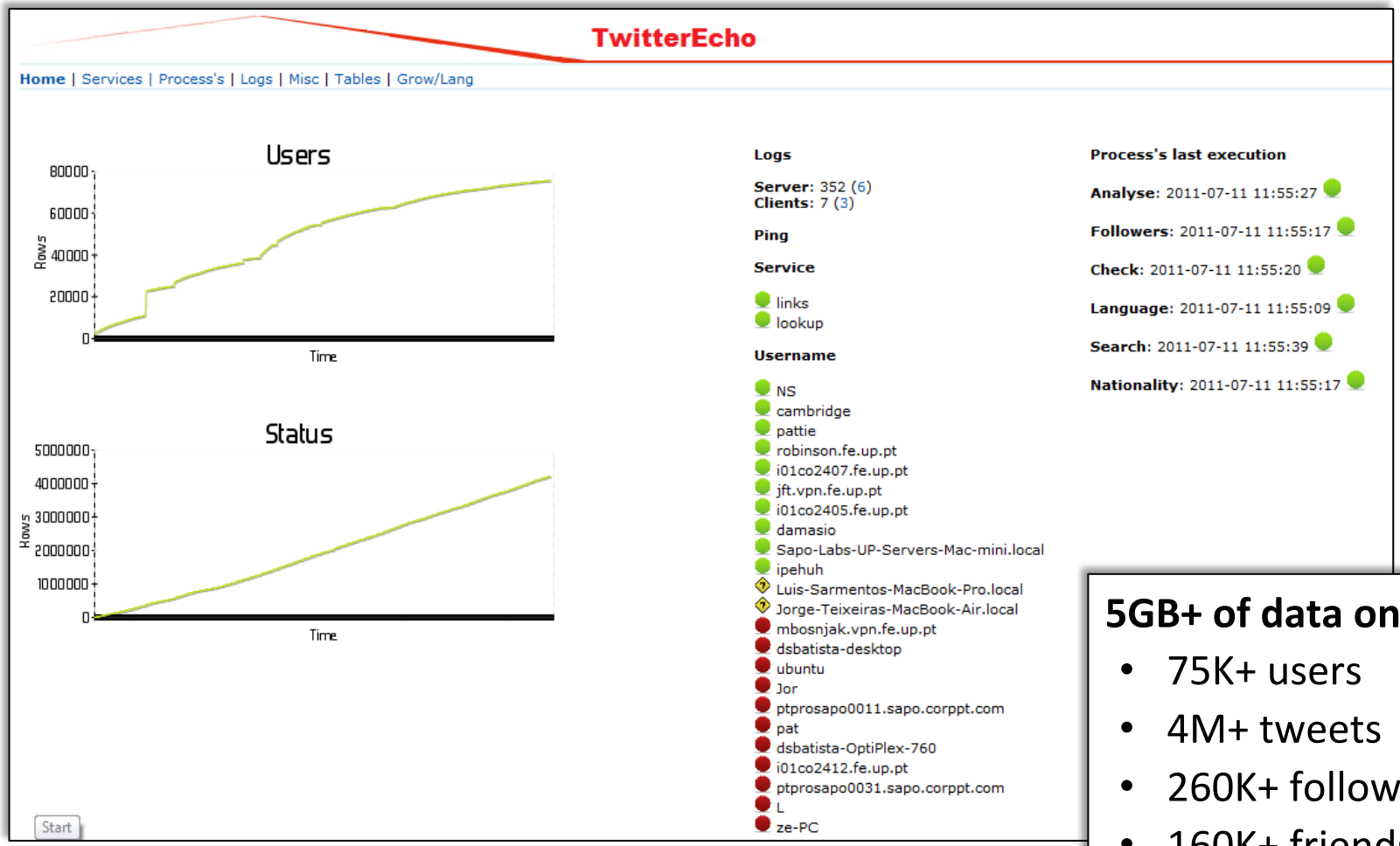
  First steps: Twitómetro

Progress Report

- 2 MSc theses completed
  - TwitterEcho crawling platform (Eduardo Oliveira)
  - Influence Analysis (José Martins)

- TREC 2011 Microblog track
  - Team: Eduarda, Gustavo, Jorge, Luis, Matko
  - IR and Graph Mining infrastructure
  - Experiments

TwitterEcho



5GB+ of data on July 10th:

- 75K+ users
- 4M+ tweets
- 260K+ followers lists
- 160K+ friends lists
- 7M+ statistics

Influence Analysis Model

- Influence indicators:
  - **Recognition** – measures the credibility and authority the community assigns to the user
    - retweet score, mention score and recommendation score
  - **Engagement** – measures the interaction with the rest of community, i.e. the user's ability to create activity around him and his content
    - number of replies received minus posted by the user
  - **Reach** – measures the size of user's audience
    - number of followers minus friends
- Comparison with other influence detection systems:

KLOUT TWITALYZER tweetgrader

Influence Analysis Baseline

Screen_name	Recognition score
Publico	100
PauloQuerido	55
itwitting	26.92
JornalNoticias	21.33
pedroaniceto	18.17
fcancio	14.95
omalestafeito	13.48
jamesdrodriguez	12.72
dita_dura	12.32
ChicodeOeiras	11.96
JMF1957	11.84
SICNoticias	11.38
supirinho	9.02
der_terrorist	8.99
cvazmarques	7.62
tvi24_iol	7.14
cmjornal	6.93
ExpressoOnline	6.87
AldaTelles	6.08
G_L	5.73

Screen_name	Engagement score
ChicodeOeiras	99.97
pulsar_ana	98.14
BUGabundo	84.67
fcancio	79.92
brunum	69.18
luis_grave	63.34
pedroaniceto	62.28
supercurlystic	62.28
jamesdrodriguez	60.53
supirinho	59.67
sophieecore	57
_Andrea_D	55.74
Wonderm00n	55.13
trainmaniac	51.66
G_L	51.15
vivi_marta	50.51
vidademigalhas	49.6
isv5	48.18
MiGoncalves7	46.2
FatimaRD	45.59

Screen_name	Reach score
cristiano	100
cacelico	4.49
R9FALCAO	2.92
zerohora	2.55
MiKe1155	2.37
simao20sabrosa	1.96
portugalnet	1.35
IgorTiago_h	1.29
katyperrynet	1.2
genevievennaji1	1.16
DiscoLee	1.11
apombalivre	1.08
jamesdrodriguez	1.05
fguarin13	1.03
webmilionario	0.9
WelitonGaspar	0.9
Corpodormente	0.78
soccerportugal	0.78
fernandoalvim	0.73
funnyhumour	0.71

Screen_name	Global influence score
Publico	80.48
PauloQuerido	49.51
ChicodeOeiras	27.56
fcancio	26.35
pedroaniceto	25.75
itwitting	21.83
jamesdrodriguez	21.1
pulsar_ana	19.42
supirinho	17.96
dita_dura	17.86
BUGabundo	17.48
JornalNoticias	17.33
omalestafeito	16.29
brunum	14.32
luis_grave	14.22
Wonderm00n	13.81
G_L	13.79
supercurlystic	13.47
vivi_marta	11.73
JMF1957	11.01

Twitter Data Curation

- Incremental extraction and representation of structural data (i.e., user and content relations)
  - Explicit social networks (followers, friends)
  - Implicit social networks (@replies, RT, mentions)
  - Hashtags co-occurrence network (folksonomy)

- Main concerns:
  - Providing chronologic navigation of data (temporal DM)
  - Efficient querying and processing of big graphs (Neo4j graph database)
  - Allowing simple and scalable access to data (easy to use API)

Data Pre-processing

- Forensic analysis
  - Identification of the author
  - Training set of ~60 messages from each of three users, we can predict the author of a message more than half the times; increasing training set to 2000 messages per author, precision can raise to 2/3 of the times
  - The most revealing feature group was the emotion group, i.e., smileys, LOLs, and so on
- Bot identification
  - Compared a number of strategies, such as posting frequency, use of URLs, the **Twitter client** used and **stylistic features**
  - Issues: little data to work with… more experiments planned
- Publications:
  - Sousa-Silva, R.; Laboreiro, G.; Sarmento, L.; Grant, T.; Oliveira, E. & Maia, B. 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages Procedings of the 16th International Conference on Applications of Natural Language to Information Systems, 2011
  - Laboreiro, G.; Sarmento, L.; Oliveira, E. Identifying automatic posting systems in microblogs Encontro Português de Inteligência Artifical, 2011

TREC 2011 – Microblog Track

- **DATA:** 2 weeks of tweets
  - circa 16M tweets
  - 24th January – 8th February
  - HTML or JSON (richer) format
- **TASK:** query at a specific time
  - Real-time search over a topic (50 topics)
  - Relevant tweets from newest to the oldest
  - Interesting new tweets favoured!

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UT Austin | Portugal
INTERNATIONAL COLLABORATORY FOR EMERGING TECHNOLOGIES, COLAB
U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

TREC 2011 – Microblog Track

- **SUBMISSIONS:** Up to 4
  - At least one without future or external data

- **EVALUATION**
  - Only English tweets
  - Return 1000 tweets, precision@30
  - Recency and interestingness

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UT Austin | Portugal
INTERNATIONAL COLLABORATORY FOR EMERGING TECHNOLOGIES, COLAB
U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Dataset Statistics

- Downloaded corpus:
  - 16.141.812 tweets (15.812.442) in JSON format (information rich)
  - 5.1M replies (32%)
  - 1.6M retweets (10%)
  - 4.6M mentions (17%)
  - 2.2M distinct URLs (16%)
  - 478k hashtags (12%)

Experiments

- Indexing service : Terrier IR Engine
- In-house tweet tokenizer
- URL expansion
- Language detection (English tweets only!)

- Hashtag analysis and tracking through time
- Explicit and implicit network creation
- Query enrichment using analyzed URLs

Exploiting the Solution

- A giant dataset to experiment with (although not as good as TwitterEcho dataset)
- We will effectively have a state-of-the art system for real-time search over (Portuguese) twittosphere
- Able to add modules
  - Our tokenizer
  - NERC
  - Retrieval based on scored influence (TBD)

Next Steps

- Participation in TREC - **DEADLINE:** 11th August
- Influence detection + topic modelling
  - Research on topic modelling over Twitter data, considering topics evolve over time
  - Further research on methods for modelling the influence, based on simple network topology and network metrics
- Integrating Twitómetro and TREC solutions in the REACTION prototype